Audio-cough event detection based on moment theory

Monge-Alvarez, Jesus; Hoyos Barceló, Carlos; Dahal, Keshav; Casaseca, Juan Pablo

Published in:
Applied Acoustics

DOI:
10.1016/j.apacoust.2018.02.001

Published: 01/06/2018

Document Version
Peer reviewed version

AUDIO-COUGH EVENT DETECTION BASED ON MOMENT THEORY

Jesús Monge-Álvarez[a,*], Carlos Hoyos-Barceló[a], Keshav Dahal[a,c], Pablo Casaseca-de-la-Higuera[a,b]

[a]Centre for Artificial Intelligence, Visual Communications, and Networks. School of Engineering and Computing. University of the West of Scotland.

[b]Laboratorio de Procesado de Imagen. ETSI Telecomunicación. Universidad de Valladolid.

[c]Nanjing University of Information Science and Technology (NUIST), China.

*Corresponding author at: School of Engineering and Computing, Paisley Campus, University of the West of Scotland, High St, Paisley, PA1 2BE, United Kingdom. Current e-mail address: jesus.monge@uws.ac.uk Permanent e-mail address: jsmonge@outlook.es

ABSTRACT

Cough detection has recently been identified as of paramount importance to fully exploit the potential of telemedicine in respiratory conditions and release some of the economic burden of respiratory care in national health systems. Current audio-based cough detection systems are either uncomfortable or not suitable for continuous patient monitoring, since the audio processing methods implemented therein fail to cope with noisy environments such as those where the acquiring device is carried in the pocket (e.g. smartphone). Moment theory has been widely applied in a number of complex problems involving image processing, computer vision, and pattern recognition. Their invariance properties and noise robustness make them especially suitable as "signature" features enabling character recognition or texture analysis. A natural extension of moment theory to one-dimensional signals is the identification of meaningful patterns in audio signals. However, to the best of our knowledge only marginal attempts have been made in this direction. This paper applies moment theory to perform cough detection in noisy audio signals. Our proposal adopts the first steps used to extract Mel frequency cepstral coefficients (time-frequency decomposition and application of a filter bank defined in the Mel scale) while the innovation is introduced in the second step, where energy patterns defined for specific temporal frames and frequency bands are characterised using moment theory. Our results show the feasibility of using moment theory to solve such a problem in a variety of noise conditions, with sensitivity and specificity values around 90%, significantly outperforming popular state-of-the-art feature sets.

KEYWORDS

Audio event detection, cough segmentation, moment theory, *k*-Nearest Neighbours, Time-frequency analysis.

1 Introduction

Cough is a symptom associated with over one hundred medical conditions like respiratory diseases (e.g., asthma, bronchiectasis, or chronic obstructive pulmonary disease), generic pathologies such as cold or allergies or even lifestyle (smokers) [1]. Respiratory conditions constitute a significant burden for national health systems and economies [2], [3], with clear potential to be released if objective continuous monitoring of symptoms such as cough was made possible. Consequently, cough detection has recently been identified as of paramount importance to fully exploit the potential of telemedicine in respiratory conditions and thus decrease their economic burden [4].

Audio cough events are non-stationary signals presenting a sparse spectrum that exhibits a high-energy peak around 400 Hz and a secondary peak between 1 and 1.5 kHz. Detecting and properly characterising them is hindered by a lack of a clear pitch structure [5]. Moreover, there exist other events produced by the human body such as throat clearing, gasping breath or laugh whose acoustic properties are very similar. Also, a continuous monitoring environment (e.g. when carrying a smartphone in the pocket) can prevent accurate detection due to the presence of noise with diverse spectral content.

Audio event detection (AED) was originally posed as a binary classification problem to differentiate speech from non-speech events. Such systems are commonly known as voice activity detectors [6]. Later, the generalisation of other information sources such as audio/video streaming, musical repositories or online video-games [7], [8], [9] led to other applications involving non-speech signals: query-by-humming, recommender systems or automatic music transcription [10]. The signal processing and machine learning techniques applied within this new context are often referred to as *machine hearing* [11]. Moreover, the emergence of new devices – e.g. smartphones, tablets or *wearables* – with their increasing computational capabilities diversified the variety of applications requiring AED [6]. These new applications were initially focused on content-based audio classification and retrieval [10], [12]. However, nowadays there is an increasing number of applications such as medical telemonitoring [13], ambient sound recognition [14], or audio surveillance (e.g. monitoring of wildlife areas [15] or classification of aircraft noise [16]).

Despite the fact that automatic detection and analysis of speech are still active research areas [17], their methods are not always directly applicable to other AED problems [6] due to two main reasons. First, speech or music repositories are in general larger than other

databases that are more difficult to record or whose production is less frequent [18]. This point often leads to suboptimal or unfeasible applications of speech/music-specific methods to these smaller-sized datasets. Second, the differences in acoustic (e.g., formant frequencies or pitch contour) and spectral (e.g. distribution or spread) properties among different audio signals play an important role. Many of these methods are specifically designed on the basis of speech properties [17]. Thus, their application to other types of audio events such as acoustic biomedical signals or environmental sounds does not always produce satisfactory results [6].

A number of papers have addressed the problem of automatic cough detection from different perspectives. Commercial cough detectors achieve sensitivity values in the 80% range by employing features extracted not only from cough sounds but also from chest movement [19], [20]. Matos *et al.* employed a keyword-spotting approach based on a hidden Markov model. Their average detection rate was 82% [21]. Drugman used mutual information-based measures and feature synchronisation to perform feature selection and classification for cough segmentation. Sensitivity and specificity values above 90% were reported [22].

Other authors have designed specific methods for cough segmentation. You *et al.* employed non-negative matrix factorisation, reaching sensitivity and specificity values around 85% [23]. They also proposed an ensemble multiple frequency subband features approach where recall values around 74% with an overall 82% performance were reported [24]. Finally, deep learning methods based on convolutional neural networks (CNN) and recurrent neural networks (RNN) have recently been used as well. Amoh and Odame achieved 83% sensitivity using this approach, although the CNN was superior (93%) in terms of specificity where the RNN only achieved 75% [25].

A number of approaches aiming at robust identification of audio events rely on interesting principles that could be adopted for cough detection. These approaches as such could only be applied to cough identification and not to cough detection, since they all work on previously segmented events of interest. Foggia *et al.* employed a *bag-of-audio-word* approach aimed at improving the discriminative power while the classification scheme is kept simple [26]. Dennis *et al.* developed spectrogram image features (SIF) for sound event classification. The spectrogram is normalised into greyscale, and its dynamic range is quantised into regions before partitioning it into blocks whose distribution statistics are extracted to build a feature set for classification. The main disadvantage of this approach is

the large dimension of the feature set (486) [27]. This drawback was partially solved by Sharan and Moir who improved the basic SIF approach for robust audio surveillance. They reduced the feature space dimension to 216 by computing the mean and standard deviation of the distribution statistics across rows and columns [28]. Other time-frequency representations have also been employed. *Cochleagram* image-based feature computation has found usage in speech recognition and audio separation applications [6]. The Wavelet transform, has also been used for speech and music discrimination since it provides better time and frequency localisation [6]. Finally, unsupervised classification approaches for environmental noise signal classification have recently been proposed [29].

Although extensively applied in image processing and computer vision, moment-based methods are still marginal in one-dimensional signal processing. These methods hold a number of features that make them suitable for cough detection due to their 2D nature. As image processing methods, they can be applied to windows including a time-frequency representation of the signal (e.g., spectrogram, *cochleagram*, as in the robust methods described above) and exploit this higher dimensionality to achieve robust cough detection. Recently, Sun *et al.* employed features based on local Hu moments (HUm) for speech emotion recognition [30]. Our previous work [13] showed that a similar approach could be successfully used to perform robust detection of audio-cough events. To the best of our knowledge, only the extensions of HUm in [30] and [13] have been applied so far to audio signal processing. The two examples described above show the promising applicability of moment theory for cough detection in particular and more generally for audio processing.

This paper proposes a novel methodology to extend moment theory to audio signal processing with a specific application to cough detection in noisy environments[1]. The individual pattern discrimination capability and robustness against noise of different moment families are studied and a discussion on the hyper parameter settings and design decisions in the methodology is presented. Our results show that using audio features based on moment theory significantly outperforms popular state-of-the-art feature sets such as Mel frequency cepstral coefficients (MFCC) [6] and linear predictive cepstral coefficients (LPCC) [30], especially in low Signal to Noise Ratio (SNR) scenarios. We also show that our method overcomes more noise-robust feature sets such as spectral subband centroid histograms (SSCH) [31] and power normalized cepstral coefficients (PNCC) [17]. It is worth

[1] We will employ the broad term "noisy" to refer to conditions in which unwanted signals overlap with the audio event of interest, regardless of their random or deterministic nature.

highlighting at this point that in our context, cough detection is understood as a continuous process carried out while the signal is recorded in a *soft* real-time manner. We do not hold the assumption that the events are pre-segmented in different classes for further automatic classification as Audio Event Identification methods require.

The paper is structured as follows: Section 2 introduces the proposed methodology, including a description of the taxonomy of the different moment families selected for the study. Section 3 presents the experimental setup, including the design of the employed cough database, and the performance measurements used to evaluate the proposal. The experimental results are presented in Section 4 and discussed in Section 5. Both sections validate the proposal by justifying the adopted methodology against previous approaches and studying its sensitivity with respect to different parameter configurations and design choices. Section 6 finalises the paper with some conclusions and future directions.

2 Proposed Methodology

2.1 Extension of moment theory to audio event detection

Many audio processing features are based on the spectral energy distribution of the acquired signals. For non-stationary signals, some type of Short Time Fourier Transform (STFT) computation provides a time-frequency decomposition to account for this spectral distribution along time. To obtain such representation, a filter bank is built to characterise the spectrum in several frequency bands. As an example, MFCC, one of the most widely used feature sets, employs the Mel frequency scale to set the limits of each filter in the filter bank. Once the filter bank is applied, the logarithm is computed for all energy values to obtain a representation close to the response of the human cochlea. This is the starting point of the proposed methodology, presented in the following paragraphs.

In order to build a time-frequency distribution, the one-sided normalised power spectral density ($PSD_k[f]$, $k=1,\ldots,K$) is first estimated for each window as the Fourier transform of the autocorrelation function according to the Wiener-Khinchin-Einstein theorem [32]. Secondly, the logarithm of the energies is computed for every window in a series of bands defined by a filter bank in the Mel scale:

$$E_k(m) = \log\left(\sum_{f=f_{\min}}^{f_{\max}} PSD_k[f] \cdot H_m[f] \right) \quad 1 \le m \le M \tag{1}$$

where m denotes each filter within the filter bank, and f_{min} and f_{max} are the minimum and maximum frequencies considered in the analysis. The filter bank is defined as:

$$H_m[f] = \begin{cases} 0, & f < C(m-1) \\ \dfrac{2 \cdot (f - C(m-1))}{(C(m+1) - C(m-1)) \cdot (C(m) - C(m-1))}, & C(m-1) \le f < C(m) \\ \dfrac{2 \cdot (C(m+1) - f)}{(C(m+1) - C(m-1)) \cdot (C(m+1) - C(m))}, & C(m) \le f \le C(m+1) \\ 0, & f > C(m+1) \end{cases} \tag{2}$$

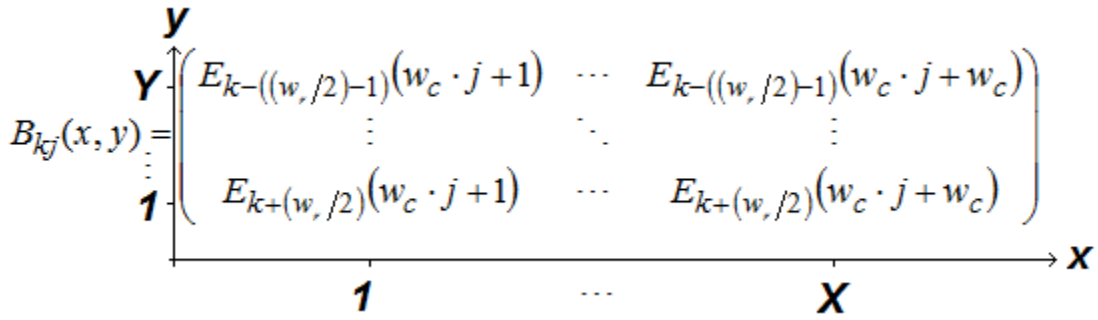
$C(m)$ $1 \le m \le M$ are the central discrete frequencies for each filter in the filter bank, uniformly spaced between f_{min} and f_{max} in the Mel scale. Conversion from linear frequency scale to the Mel is performed as:

$$f[Mel] = 2595 \cdot \log_{10}(1 + f[Hz]/700) \tag{3}$$

Consequently, after performing the first step for all the windows, a $K \times M$ matrix, E, is obtained, with K the number of windows and M the total number of filters defined in the filter bank. Fig. 1 illustrates how E is obtained.

For short windows in the STFT decomposition, several adjacent windows may belong to the same audio event. Similarly, the energy in w_c adjacent filters can be joined together with w_r consecutive windows resulting in ($w_r \times w_c$) sub-matrices within the energy matrix E. These sub-matrices, which hold energy patterns of a particular temporal frame and for a given frequency band can, therefore, be characterised by calculating their moments. The rationale of using moments to infer spectral signatures from energy patterns in the spectrogram resides in their ability to characterise spatial patterns in images. As stated in the introduction, moments are image features widely used for different tasks – for instance, digit reconstruction [33] or digit recognition [34], [35]. Digits are small images (or small parts of a larger image). Moments are able to identify spatial patterns so that different digits can be recognised. Therefore, moment theory constitutes a good candidate to extract meaningful features form spectral energy patterns in the time-frequency domain.

From the energy matrix E, the computation of the final feature space is carried out as follows: Energy patterns with ($w_r \times w_c$) size are first built by dividing E into blocks B_{kj}. For the k-th window, these blocks are constructed as:



$$B_{kj}(x,y) = \begin{pmatrix} E_{k-((w_r/2)-1)}(w_c \cdot j + 1) & \cdots & E_{k-((w_r/2)-1)}(w_c \cdot j + w_c) \\ \vdots & \ddots & \vdots \\ E_{k+(w_r/2)}(w_c \cdot j + 1) & \cdots & E_{k+(w_r/2)}(w_c \cdot j + w_c) \end{pmatrix}$$

$$j = 0,\ldots,(M/w_c)-1 \tag{4}$$

In Equation (4), x and y respectively refer to the indexing variables in the horizontal and vertical axes, respectively. X and Y are the number of elements in the horizontal and vertical axes.

The blocks for the first and last windows are padded with zeros on the top or bottom, respectively, up to the (w_r × w_c) size where no more data from the E matrix are available. This padding has a negligible effect on the calculations since the method is thought to perform audio-cough detection in a *soft* real-time manner, i.e., long audio recordings are processed using overlapping windows. The windows requiring padding encompass a maximum time period of hundreds of milliseconds, so padding does not have a significant impact on performance. At the end of block construction step, each window is represented by D energy patterns:

$$D = M/w_c \tag{5}$$

The selected moments described in the next sections are calculated for each of the energy patterns B_{kj}. After computing these moments for all windows, a D-dimensional feature vector is obtained for each of the K windows. The magnitude is computed for complex moments. The process of building energy patterns and the generation of the final feature space is summarised in Fig. 2 for the sake of clarity. The steps shown in Fig. 2 constitute the main contributions of the proposed work.

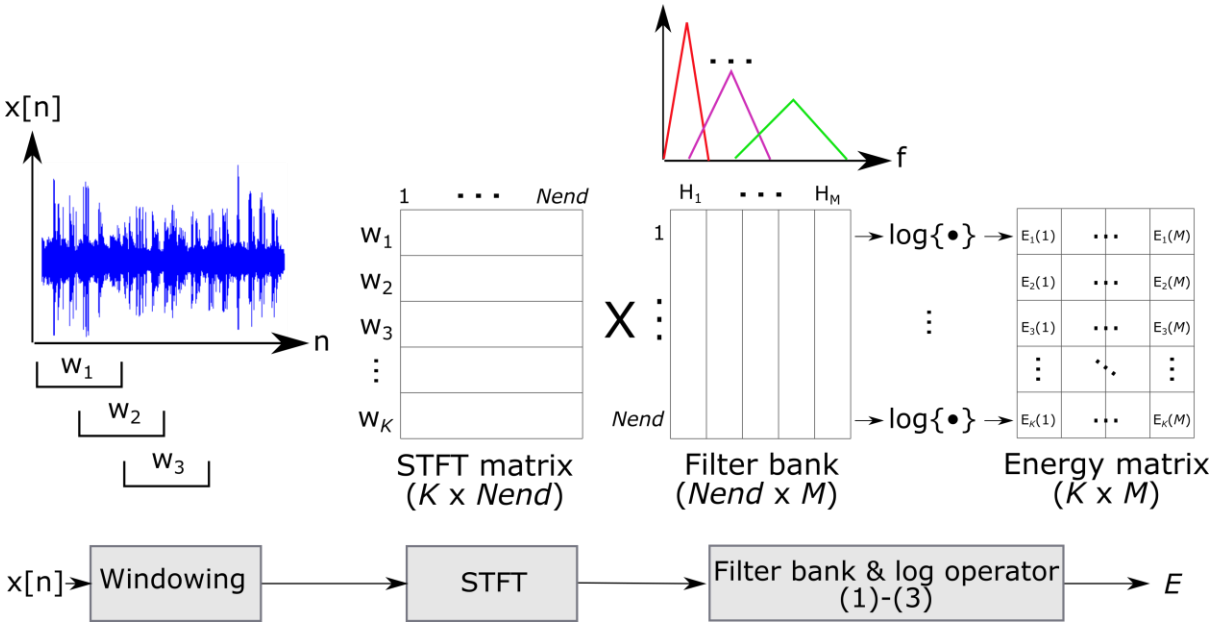


Fig. 1 Pipeline of the first step of our proposal. This processing follows the same approach as other audio feature calculation such as MFCC. *Nend* refers to the number of points of the one-sided version of the STFT.

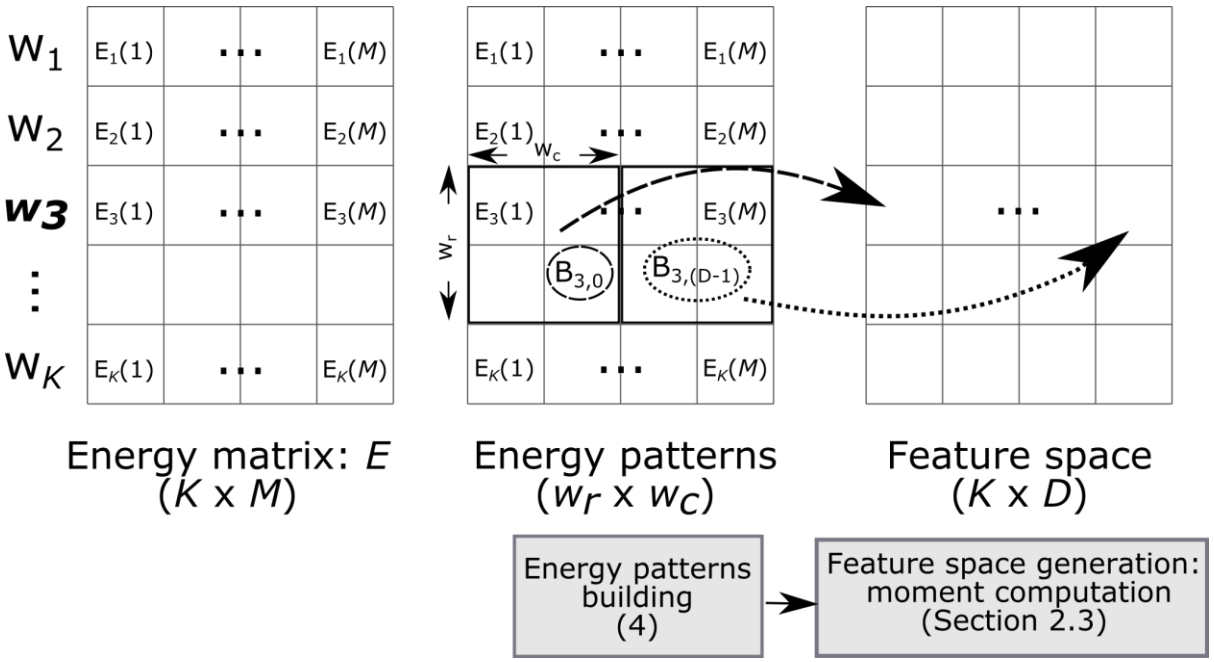


Fig. 2 Pipeline of the second step of our proposal. After building energy patterns for every window (an example is depicted for the third window and a (2 × 2) block size) moments are used to characterise them, generating the final feature space.

Fig. 2 also shows an interesting property of the proposed method. The dimension of the feature space is determined by the number of filters and the width of the energy patterns in the frequency axis (see equation (5)).

These two parameters (M and w_c) can be adjusted to set the dimension of the feature space. If large dimensions do not constitute a problem, the number of filters can be increased,

or w_c can be reduced. However, both choices have their counterparts. Regarding the first, the filters are overlapped, since their limits have been defined in the Mel scale. A high number of filters may cause the information redundancy due to overlapping to spoil the benefit of a more fine-grained spectral characterisation. As for the second option, the user should take into account the feasibility of defining moment polynomials for small block sizes (see Section 3.2). In general, for low order polynomials, this does not constitute a problem, as they also have few zero-crossings. On the other hand, if the number of zero-crossings is higher than (w_c - 1), the definition of the polynomial in that axis is not feasible (see Section 2.3).

Considering the size of the block in the time axis, w_r, the same considerations are valid for the lower limit. As for the upper limit, the following applies: For windows located in the middle of an event, it is highly likely that adjacent windows also belong to the same event. Nevertheless, for those closer to the boundary between two events, some of them may belong to a different event. The larger the value of w_r, the higher the number of windows that may belong to a different event when building the blocks in the boundary.

2.2 Selected moment families

Complex moments are the most generic moment family [36], encompassing the rest of families, namely: geometric, rotational and orthogonal moments [37]. The basis of geometric moments is the theory of algebraic invariants [38]. Their main advantages are that they are computationally simple and they can be physically interpreted at lower orders. On the other hand, they suffer from a high degree of information redundancy, since the employed basis for projection is not orthogonal, and the higher the order is, the more noise-sensitive they are. Additionally, there is a large variation in their dynamic range of values [37]. As such, geometric moments do not have any invariance property, although using them together with central moments, Hu achieved the definition of three moments which are invariant to rotation, scaling, and translation, the so-called HUm [39]. Hu's work could be considered the pioneering work in moment theory. To overcome the shortcomings of geometric moments, Teague proposed the use of orthogonal moments [40], i.e. their kernels are orthogonal polynomials. Their orthogonality allows for better image representation and reconstruction – derived from their lower degree of information redundancy – with the added value of high noise robustness [37].

Orthogonal polynomials can be defined using continuous or discrete functions. The major disadvantages of continuous moments compared to discrete ones are that they require a transformation of the coordinate space as well as a suitable transformation of the involved integral calculations, which lead to discretisation errors. Besides, discretisation errors accumulate as the moment order increases, and this limits the accuracy of the computations. Finally, similar to geometric moments, they present a considerable variation in their dynamic range of values. Some examples of continuous orthogonal moments are Lm [40] or FMm [35]. Among the discrete group, Tm [41], Km [42] or dHm [43] are some of the most representative examples.

Together with orthogonality and type of variable (continuous or discrete), additional properties need to be considered for some moment families. For instance, FMm are defined in polar coordinates whereas Lm use Cartesian coordinates. The remaining orthogonal moments (all discrete) also use Cartesian coordinates. However, Km and dHm have additional properties in comparison to Tm, the first discrete orthogonal moment proposed by Teague [40]. Km can extract local features from any region of interest of an image. In other words, the Krawtchouk polynomials (Kp) employed in their calculation can be located in a particular position to emphasise the characterisation of that area. We have referred to this property as "locality". The parameters that control locality are p_x and p_y [42] (see next section).

In the case of dHm, locality is determined by two parameters, namely a and c [43] (see next section). Likewise, dHm have an extra property compared to Tm and Km: they are defined in a non-uniform lattice. This means that Tm and Km are directly defined on the image grid but, for dHm, an intermediate step must be introduced to get the non-uniform lattice, $x(s) = s \cdot (s + 1)$. The main properties of each selected moment are summarised in Table 1.

We have selected HUm in this work for two reasons: (a) they were the first proposal in the moment theory, so they became a *de facto* standard against which to compare the performance of new proposals; (b) to the best of our knowledge, no other type of moments has been used for both image [37], and audio signal processing [13].

Regarding continuous moments, we selected FMm and Lm since they are defined in different coordinate systems. Comparing both of them will provide insight on which coordinate system is more suitable when using moments for audio processing. By comparing

these continuous moments with discrete ones, we can check whether the disadvantages of the former when they are used for image processing are also present in this new context.

The choice of discrete orthogonal moments (Tm, Km, dHm) was based on the additional properties that these moments offer. We included Tm as the most basic example of this family while Km and dHm will be used to analyse the contribution of the properties of locality and a non-uniform lattice, respectively. Using this set of six moments we expect to explore the behaviour of some of the essential moment properties in our attempt to extend their use for audio processing.

Moment	Type of variable	Orthogonality	Coordinate system	Locality	Type of lattice
HUm	Continuous	No	Cartesian	No	Uniform
FMm	Continuous	Yes	Polar	No	Uniform
Lm	Continuous	Yes	Cartesian	No	Uniform
Tm	Discrete	Yes	Cartesian	No	Uniform
Km	Discrete	Yes	Cartesian	Yes	Uniform
dHm	Discrete	Yes	Cartesian	Yes	Non Uniform

Table 1 Properties of the selected moments

The computation of the selected moments using the $X \times Y$ energy patterns $B_{kj}(x,y)$ as input, is described in the following sub-sections.

1) Hu moments (HUm)

Hu defined three invariant moments as [39]:

$$HUm(order=1)=\eta(p=2,q=0)+\eta(p=0,q=2) \tag{6}$$

$$HUm(2)=(\eta(2,0)-\eta(0,2))^2+4\cdot(\eta(1,1))^2 \tag{7}$$

$$HUm(3)=(\eta(3,0)-3\cdot\eta(1,2))^2+(3\cdot\eta(2,1)-\eta(0,3))^2 \tag{8}$$

where

$$h(p,q)=\frac{m(p,q)}{m(0,0)^g} \text{ where } \gamma=\frac{p+q}{2}+1 \tag{9}$$

$$\mu(p,.q)=\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}(x-\bar{x})^p\cdot(y-\bar{y})^q\cdot B_{kj}(x,y) \tag{10}$$

$$\psi(p,q)=\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}x^p\cdot y^q\cdot B_{kj}(x,y) \tag{11}$$

where *p, q,* are the raw moments' order in their respective dimension, $\gamma(p,q)$ and $\mu(p,q)$ are the non-central and central moments respectively calculated over the energy blocks, and $\eta(p,q)$ is the normalised central moment. In equation (10), $\bar{x}=\psi(1,0)/\psi(0,0)$ and $\bar{y}=\psi(0,1)/\psi(0,0)$ are normalised first order non-central moments. Equations (6-8) are used to obtain the invariant HUm from Equations (9-11). We refer to [39] for a more detailed description of these parameters.

2) Fourier-Mellin moments (FMm)

The FMm are defined in the polar coordinate system (r, θ) with $0 \le r \le 1$, so the first step is to map the pixel grid into polar coordinates. Once the (r, θ) values for each pixel are known, the Fourier-Mellin polynomials (FMp) are defined as follows [35]:

$$FMp(p,r)=\sum_{k=0}^{p}(-1)^{p+1}\cdot\frac{(p+k+1)!}{(p-k)!\cdot k!\cdot(k+1)!}\cdot r^{k} \tag{12}$$

The FMm of the image can be expressed in axial coordinates (x, y) as follows:

$$FMm(p,q)=\frac{p+1}{\pi}\cdot\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}B_{kj}(x,y)\cdot FMp(p,r)\cdot e^{-j\cdot q\cdot\theta} \tag{13}$$

In equations (12) and (13), p is the order of the FMp and the FMm, with $p \ge 0$, and q is the repetition, with $q = 0, \pm1, \pm2, \ldots$.

3) Legendre moments (Lm)

The set of Legendre polynomials (Lp) forms a complete orthogonal basis in the interval [-1, 1]. The $X \times Y$ pixel grid must be normalised to $\begin{pmatrix}-1,\ 1\\ -1,\ 1\end{pmatrix}$. The discrete approximation of Lm is computed as [44]:

$$Lm(p,q)=A\cdot\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}Lp(p,x_n)\cdot Lp(q,y_n)\cdot B_{kj}(x,y) \tag{14}$$

In equation (14), $A=(2\cdot p+1)\cdot(2\cdot q+1)/X\cdot Y$ and p and q are the order of the polynomials defined in the x and y axes, respectively. The final order of the Lm is (p + q). $Lp(p,x_n)$ can be recursively computed according to equations (15), (16) and (17):

$$Lp(0,x_n)=1 \tag{15}$$

$$Lp(1,x_n)=x_n \tag{16}$$

$$Lp(t+1,x_n)=\frac{2\cdot p+1}{p+1}\cdot x_n\cdot Lp(t,x_n)-\frac{p}{p+1}\cdot Lp(t-1,x_n) \quad (17)$$

Equations (15)-(17) apply similarly for $Lp(q,y_n)$.

4) *Tchebichef moments* (Tm)

The (p + q)-order Tm are computed as follows [41]:

$$Tm(p,q)=\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}Tp(p,x)\cdot Tp(q,y)\cdot B_{kj}(x,y) \quad (18)$$

A recursive formula to compute the Tchebichef polynomials (Tp) is provided in equations (19), (20) and (21) [45]:

$$Tp(0,x)=1 \quad (19)$$

$$Tp(1,x)=\frac{2\cdot d+1-X}{X} \quad (20)$$

$$t\cdot Tp(t,x)=(2\cdot t-1)\cdot Tp(1,x)\cdot Tp(t-1,x)-(t-1)\left(1-\frac{(t-1)^2}{X^2}\right)\cdot Tp(t-2,x) \quad (21)$$

Equations (19)-(21) apply similarly to $Tp(q,y)$.

5) *Krawtchouk moments* (Km)

The (p + q)-order Km expressed in terms of Krawtchouk polynomials (Kp) are defined as [42]:

$$Km(p,q)=\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}Kp(p,p_x,x,X-1)\cdot Kp(q,p_y,y,Y-1)\cdot B_{kj}(x,y) \quad (22)$$

A recursive algorithm to compute the Kp is provided in [45]. In this work, the value of p_x and p_y is 0.4.

6) Dual Hahn moments (dHm)

Finally, the (p + q)-order dHm based on dual Hahn polynomials (dHp) are computed as follows [43]:

$$dHm(p,q)=\sum_{s=a}^{b-1}\sum_{t=a}^{b-1}dHp^{(c)}(p,s,a,b)\cdot dHp^{(c)}(q,t,a,b)\cdot B_{kj}(s,t) \quad (23)$$

where a, b and c are inner parameters. In the original work [43], the authors provide pseudo-code to compute dHp that takes advantage of the recursive properties of these polynomials, both in terms of the order (p or q) and argument (s or t). In the present study, $a = c = 0$ and $b = (a + X)$ or $b = (a + Y)$, respectively.

3 Experimental set up

3.1 Signal database

In order to assess the performance of our proposal in a variety of noisy conditions, we designed an audio signal database including a wide range of real foreground event sounds that were artificially contaminated by overlapping sounds from different environments. The following paragraphs describe the process we carried out to create the database:

1) The raw foreground events and background sounds were collected separately. We used both publicly-available audio signal databases [46] and signals recorded by ourselves. All the acquired signals were in *wav* format, at 44.1 kHz of sampling frequency and using 16 bits per sample. A Samsung S6 Edge smartphone was used to acquire the recordings. Some examples of recorded signals are speech, laugh or throat clearing. The audio-signal databases [46] provided the noisy environmental sounds that were used to contaminate the foreground events. The duration of the signals changed between few seconds and several minutes.

2) Due to the diversity of sources from the raw sounds, all the signals were normalised to have the same average power.

3) After that, we combined the audio events with the background sounds using three representative SNR values for high, moderate and low noise respectively: -6, 3 and 15 dB. To do so, we firstly selected the foreground events and the background sounds that would compose each final signal. The foreground events were collated one after the others in a larger signal. Between each foreground event, zero samples with random duration between 0.25 and 1 s were inserted. These gaps were included since two foreground events of different nature are very unlikely to occur one immediately after the other. Next, we calculated the gain factor, G, to be applied to the background sounds to achieve the desired SNR as in equation (24). Finally, both foreground events and the

background sounds were added. Fig. 3 shows the three SNR versions for one of the synthesised signals.

$$SNR_{dB} = 10 \cdot \log_{10}(1/G) \Rightarrow G = 10^{-\frac{SNR_{dB}}{10}} \quad (24)$$

As far as possible, we tried to define each signal with the greatest realism. For instance, one of the samples replicated a situation in which a jogger is practising in a park. The background sounds include her steps, wind, etc. As for the foreground events, they come up in the following order: normal breathing, sounds of breathless breathing, a cough episode and finally a throat clearing event. Neither any foreground event nor background sounds was used more than once in the synthesis. Background sounds cover both indoor (air conditioning, an office, the subway, a supermarket, toilets, a crowded restaurant, the indoor of an airport, a classroom during a lecture, a train station waiting area, a buffet restaurant, a casino, a court house, a post office, a museum, the corridor of a hospital, etc.) and outdoor (breeze, strong wind, rain under an umbrella, a crowded street, a park with children playing, a quiet residential area, a street with traffic, an open-air market, etc.) environments. Among the non-cough foreground events, the database includes throat clearing, sniffing, sneezing, burping, breathing, breathless breathing, laughs (male and female), speech (male and female), blowing nose, snoring or swallowing.

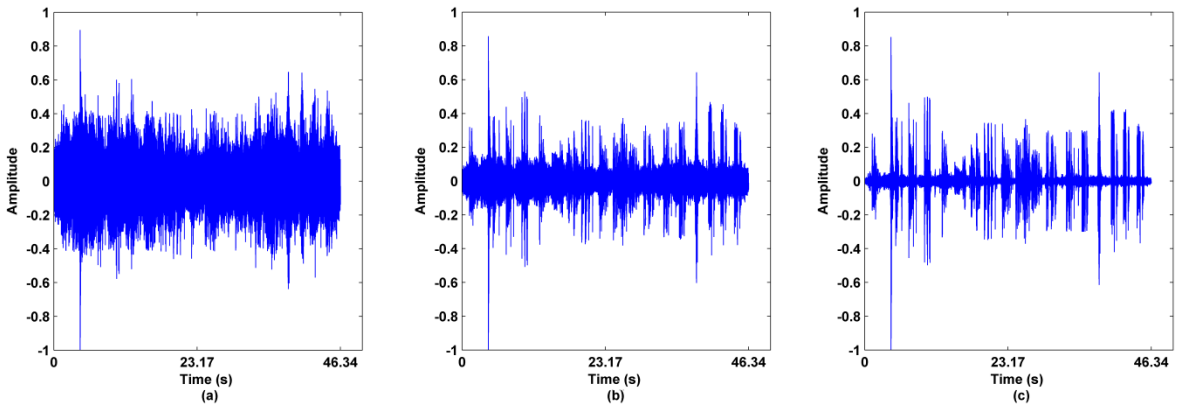


Fig. 3 Representation of one of the audio signals in the database: (a) -6 dB, (b) 3 dB and (c) 15 dB

3.2 Detection of audio-cough events

The aim of the system is to discriminate between audio-cough events and non-cough events regardless the superimposed noisy background sounds. This is posed as a two-class

pattern classification problem, where cough is the positive class any non-cough sound belongs to the negative one.

We used the following values for the block size ($w_r \times w_c$): (4 × 4), (4 × 8), (8 × 4) and (8 × 8). This allows the study of the most meaningful configuration of the blocks – i.e. symmetric in both axis with two different values, and non-simmetric with larger time or frequency axis. We have selected these block sizes since they encompass a timeframe of 125 ms ($w_r = 4$) and 225 ms ($w_r = 8$), respectively. Considering the average duration of an audio-cough event (approximately 270 ms), longer temporal gaps would increase the probability that the system misclassifies an isolated cough event. Shorter ones would not sufficiently represent the signal energy on the time scale.

The final feature dimension was set to 13 for two reasons: (a) it is a manageable dimension common in many *machine hearing* problems [13], [17]; (b) according to equation (5), the number of filters is the free parameter of our analysis; For values of $w_c = 4$ or $w_c = 8$, the number of filters is $M = 52$ and $M = 104$, respectively.

We only use the first three moment orders (0, 1 and 2 or 1, 2 and 3 depending on the moment) on the following basis:

1) To limit the initial dimension of the feature space (see Table 2).
2) In image processing, lower orders are more robust against noise than higher order ones [47]. This reason is based on the fact that moments are transparent to data semantics, so there is no difference between a pixel contaminated with additive noise and frequency bin which represents the spectrum of a foreground event plus noisy background sounds (the Fourier transform is a linear operator). Therefore, we hold the hypothesis that this property will be transferred to audio processing using the proposed approach.

As shown in Table 2, the concatenation of the first three orders results in different final dimensions, since every moment includes different number of combinations of its inner parameters within each order. To make a fair comparison of the moment families, the same final dimension should be used for all the feature sets. To do so, we employed Relieff, a widely used feature selection algorithm for binary classification [48]. The frequency band between 0 and 2000 Hz has proved enough to detect audio-cough events [13], [49] so we downsampled the database using a factor of 5, resulting in a sampling frequency of 8820 Hz. As for the remaining parameters for feature computation, the window length was 50 ms with a 25 ms shift. Finally, we employed a Kaiser window with $\beta = 3.5$ for PSD computation.

The final classification relies on a *k*-Nearest Neighbours (*k*-NN), a widely used classifier in audio signal processing [50]. The 13 selected features for each window constitute the input patterns to the *k*-NN classifier.

The classification is based on a train-validation-test partition of the feature space. 60% of the observations were used for training, 10% for validation and 30% for testing. We evaluated all the combinations among the following *k*-NN parameter values: $k=\{1, 3, 5\}$; standardised Euclidean distance, cosine distance and correlation distance. All the analysed combinations used the inverse of the distance as the weighting function, and the distances were exhaustively computed (i.e., when a new observation is to be classified the complete training space is searched to find the *k* nearest neighbours). We trained different versions of the *k*-NN classifier based on the parameter values mentioned above. Then, we evaluated their performance using the validation group. First, we selected those configurations with the highest sensitivity and, among them, the one with the highest specificity. All the moment families reported the same best configuration: *k* was set to 3 (we used this value also for Relieff algorithm), and the standardised Euclidean distance was selected.

It must be pointed out that the same partition was applied to all moments. This way, the possible differences in performance are due to the moment family and not to the feature selection or classification process. Moreover, the percentage (18.57%) of positive class observations (i.e., the percentage of windows which belong to cough events) was maintained in the three partition groups, and we used it to create a cost matrix for the 3-NN classifier. This matrix serves as a cost-sensitive parameter to deal with this unbalance in the classification step (see Fig. 4):

$$\begin{bmatrix} 0 & 1 \\ 1//0.1857 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 5.38 & 0 \end{bmatrix} \tag{25}$$

The cost matrix in equation (25) gives higher priority to false negatives as the worst case scenario for misclassification, since a window with cough events would be missed. Fig. 4 shows the pipeline of the designed system for robust detection of audio-cough events.

Moment	*p*	*q*	Order	Dimension	Final Dimension
HUm	--	--	1	13	39
	--	--	2	13	
	--	--	3	13	
FMm	0	0	0	13	39
	1	0	1	13	
	2	0	2	13	

Lm, Tm, Km and dHm	0	0	0	13	78
	1	0	1	13	
	0	1		13	
	1	1	2	13	
	2	0		13	
	0	2		13	

Table 2 Orders and dimensions of the selected moments

3.3 Figures of merit

We use sensitivity (SEN) and specificity (SPE) as basic performance measures:

$$SEN = TP/(TP + FN) \tag{26}$$

$$SPE = TN/(TN + FP) \tag{27}$$

where TP, TN, FP and FN are the number of true positives, true negatives, false positives and false negatives, respectively.

We have also defined a specific measure accounting for the robustness of the system against noise (*Noise Robustness*, NR). The NR value is derived from the accuracy ($ACC = (TP + TN)/(TP + TN + FP + FN)$) using the absolute difference of the accuracies at -6 and 15 dB:

$$NR = |ACC[15dB] - ACC[-6dB]| \tag{28}$$

Equation (28) provides a measure of the sensitivity of a specific moment with respect to SNR. A small NR means a high capability of extracting pattern information regardless the ambient noise. The NR measure must be carefully understood, since two low ACC values close to each other would result in a better NR result than two higher ACC values with a greater difference. To avoid this misunderstanding, the NR results must always be considered together with an overall *Pattern Recognition Capability* (PRC) measure. The PRC is derived from the area under the receiver operator characteristic curve (AUC). It is defined as the grand-averaged AUC across SNR and block sizes:

$$PRC = \frac{1}{3}\times\sum_{i=1}^{3} PRC_{SNR_i} = \frac{1}{3}\sum_{i=1}^{3}\frac{1}{4}\sum_{j=1}^{4} AUC[SNR\{i\}, Block\{j\}] \tag{29}$$

$SNR = \{-6,\ \ 3,\ \ 15\}$ dB and $Block = \{(4\times4),\ \ (4\times8),\ \ (8\times4),\ \ (8\times8)\}$.

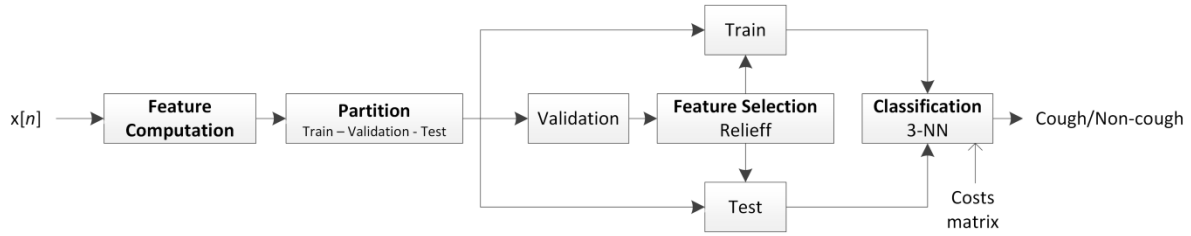


Fig. 4 System pipeline for the detection of audio-cough events

4 Results

4.1 Primary results

Fig. 5 summarises SEN and SPE results for the studied moments. The best SPE is achieved using Tm, with block size (8 × 8) and SNR = 15 dB. The best SEN is offered by the same moment and block size for SNR = -6 dB. The lowest SEN is achieved by FMm, when the block size is (4 × 8) and at SNR = -6 dB. As for the lowest SPE, it is also obtained with FMm at SNR = -6 dB and a (4 × 4) block size.

The average ACC across all SNR versions (see Table 3) shows that performance is better for w_r = 8 for w_r = 4, for all the analysed moments. Moreover, Tm offers the highest PRC value whereas the best NR is obtained for dHm when the block size is (8 × 4).

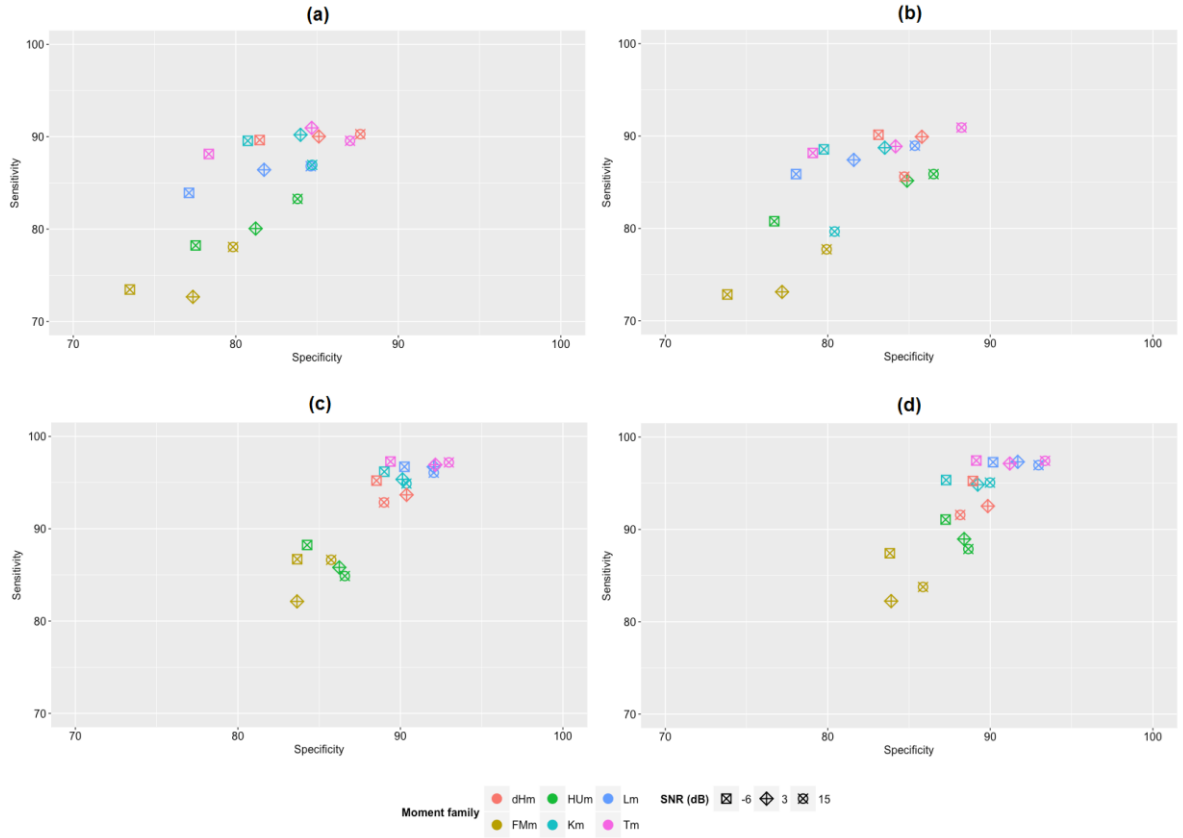


Fig. 5 SEN (%) and SPE (%) results for every block size, moment family and SNR when the feature space dimension is 13: (a) (4 x 4), (b) (4 x 8), (c) (8 x 4), (d) (8 x 8).

	NR 4x4	NR 4x8	NR 8x4	NR 8x8	ACC 4x4	ACC 4x8	ACC 8x4	ACC 8x8	AUC -6 dB	AUC 3 dB	AUC 15 dB	PRC
FMm	6.02	5.89	1.13	0.98	76.49	76.53	84.31	84.51	79.03	79.40	81.82	80.08
Lm	6.63	6.52	1.34	2.22	82.00	82.73	92.38	92.64	87.42	89.37	90.47	89.08
HUm	6.04	8.94	1.26	0.55	80.78	82.93	85.80	88.32	83.01	85.09	85.93	84.67
Tm	7.34	7.96	2.90	3.44	84.50	84.84	92.55	92.36	88.37	90.76	92.09	**90.41**
Km	2.75	1.11	0.87	2.14	84.21	82.05	90.88	89.99	88.30	89.50	87.76	88.52
dHm	5.17	0.46	**0.06**	1.33	85.71	85.28	90.15	89.74	89.03	89.66	88.73	89.31

Table 3. NR (%), average ACC (%) per block size, average AUC (%) per SNR, and PRC (%) results for feature space dimension 13. Best NR and PRC results are highlighted in boldface.

4.2 Impact of feature space dimension

The impact of w_c cannot be assessed from the preceding analysis. The number of filters duplicates when w_c moves from 4 to 8. However, the frequency range is kept, so filters are narrower. The consequence is that each energy pattern covers approximately the same frequencies regardless w_c, the only difference resides in the granularity of the energy sampling. Results in Table 3 confirm this behaviour: When comparing (4 × 4) vs (4 × 8) and

(8 × 4) vs (8 × 8) results, it can be seen that the improvement when changing w_c from 4 to 8 is not remarkable.

Therefore, a second trial of experiments was performed where the frequencies associated to each B_{kj} block were changed. M = 56 filters were used with w_c = 8, thus making the final dimension for each moment order equal to 7. After that, 7 features, instead of 13, are selected by the Relieff algorithm to build the new feature space. We only used w_c = 8 with the above configuration since it provides a coarser-grained spectral characterisation but relatively similar to (4 × 4) block size with M = 52, which is the basic configuration of the proposal. This allows: (1) to study how equation (5) affects the performance of our proposal; and equation (2) to assess whether the superiority of choosing w_r = 8 is also observed. These results are presented in Fig. 6.

By comparing Fig. 5 and Fig. 6, a performance reduction can be observed for all moments and both block sizes. PRC results in Table 4 confirm this behaviour. For example, Tm perform best, but the obtained PRC has been lowered by 6.26%. Similarly, generally speaking, PRC values have decreased for all moments. Besides, nine out of the twelve combinations exhibit worse NR. Consequently, both PRC and NR are negatively affected when the feature space dimension is reduced. Moreover, the average ACC across SNR versions (see Table 4) shows that results for (8 × 8) block sizes are better than those obtained for (4 × 8). This behaviour was also observed in the primary results.

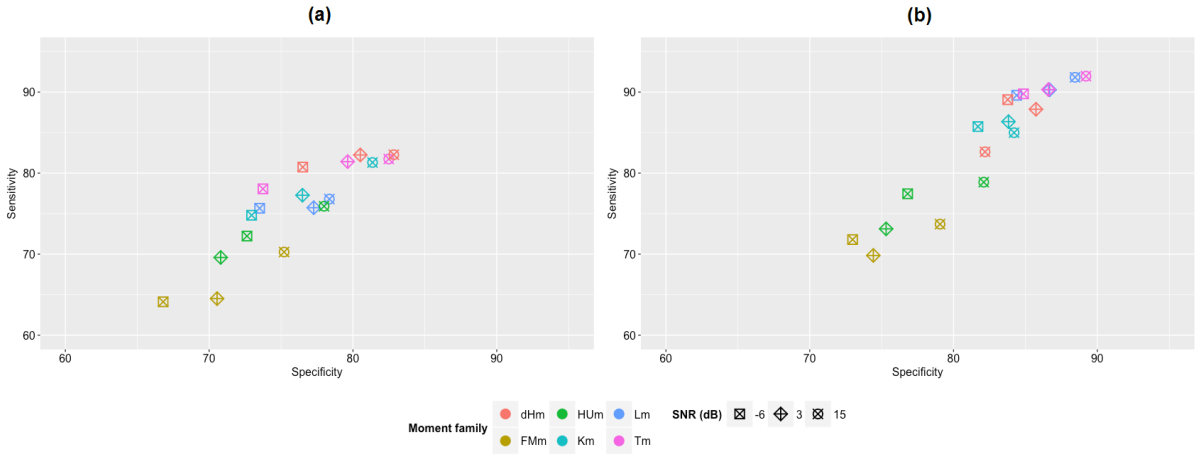


Fig. 6 SEN (%) and SPE (%) results for every block size, moment family and SNR when the feature space dimension is 7: (a) (4 x 8) and (b) (8 x 8).

	NR 4x8	NR 8x8	ACC 4x8	ACC 8x8	AUC -6 dB	AUC 3 dB	AUC 15 dB	PRC
FMm	7.99	5.31	70.00	74.80	68.92	69.83	74.56	71.10

Lm	4.16	3.73	76.32	87.26	80.79	82.49	83.86	82.37
HUm	5.04	4.57	73.58	77.77	74.78	72.20	78.72	75.23
Tm	7.81	3.95	78.95	87.59	81.61	84.49	86.35	**84.15**
Km	8.05	**1.91**	77.09	83.77	78.80	80.98	82.97	80.91
dHm	5.43	2.49	80.28	84.38	82.52	84.09	82.48	83.03

Table 4. NR (%), average ACC (%) per block size, average AUC (%) per SNR, and PRC (%) results for feature space dimension 13. Best NR and PRC results are highlighted in boldface.

4.3 Comparison to other methods

We compared our proposal with the baseline parameter configuration – (4 x 4) block size – with other methods commonly used for audio processing, namely: MFCC, LPCC, PNCC and SSCH. To perform a fair comparison, we employed the same relevant parameters for the calculations namely, 50 ms windows with 25 ms shift and a Kaiser window with $\beta = 3.5$. We used 52 filters to compute MFCC, PNCC and SSCH. LPCC are directly derived from linear predictive coefficients. The final dimension is 13. Regarding PNCC, we employed the standard parameter configuration provided by the authors in [17], whereas SSCH computation is based on histograms using 38 bins and filters with 3 Barks width [31].

Fig. 7 shows the comparison results. We have summarised all the results obtained for the different moment families in boxplots to provide a general comparison with the state-of the-art methods. All the evaluated moments outperform the state-of-the-art methods at their corresponding SNR values. All moment families yield higher average SEN and SPE results regardless the SNR. Among the four compared methods, PNCC offers the best sensitivity regardless the SNR whereas MFCC performs equivalently in terms of SPE. LPCC are the worst performing for both SEN and SPE regardless the SNR.

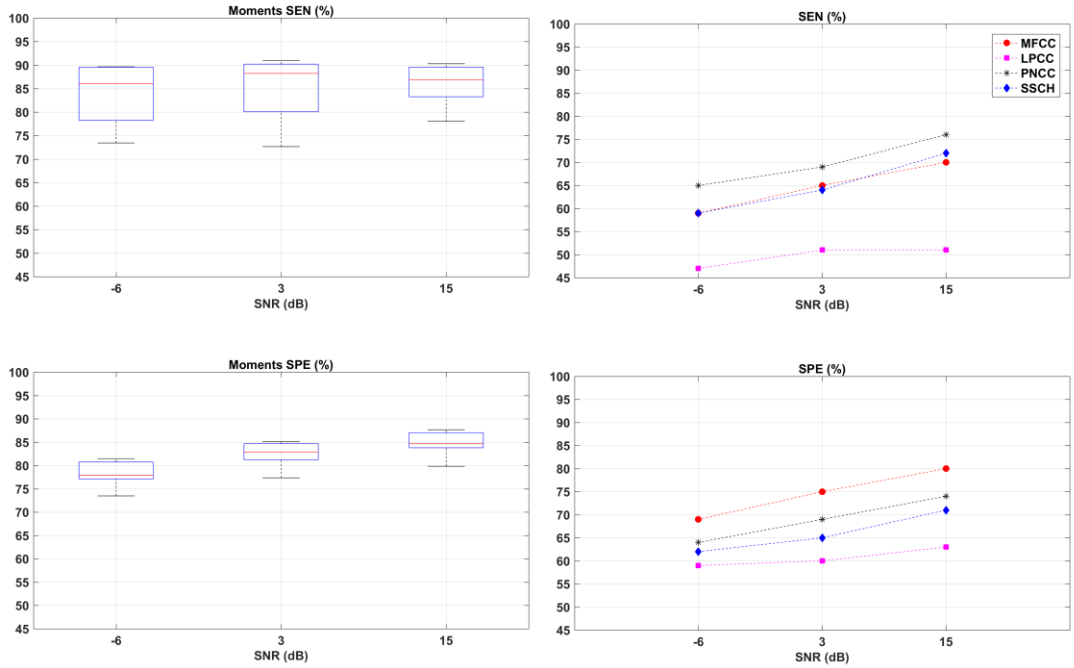


Fig. 7 Comparative results between our proposal and other methods (MFCC, LPCC, PNCC and SSCH). Boxplots in the left hand side summarise the results obtained with all the moment families. Performance for compared methods is presented on the right hand side.

5 Discussion

5.1 Analysis of the proposed method

From a holistic perspective, the obtained results are satisfactory performance-wise regardless of the moment family and the employed block size. To some extent, this validates our proposal to extend the applicability of moment theory for audio processing. Despite these good overall results, each moment family offers a different performance. Discrete orthogonal moments (Tm, Km and dHm) achieve better PRC than FMm and HUm, whereas Lm are equivalent. FMm and HUm have other properties that account for these results. FMm require a transformation of the coordinate space whereas HUm are not orthogonal and this increases information redundancy among the involved orders. Thus, the discretisation error seems to have a smaller effect than the coordinate system and, especially, the lack of orthogonality. In contrast, the original dimension in FMm and HUm is smaller (Table 2), so their computational load will be smaller as well (for FMm, we consider the transformation to polar coordinates negligible in terms of computational complexity). They are therefore less effective but more efficient.

Interestingly, Tm performed the best among discrete orthogonal moments regarding PRC but worse than Km and dHm concerning NR. Even though Tm are the simplest ones, they present two suitable properties for pattern recognition, namely: being defined in a discrete domain and orthogonality. The definition of Tm is the one that better maps to the definition of the studied energy patterns. Km and dHm allow for a better pattern analysis on the basis of their locality and the use of a non-uniform lattice, which may be the reason behind their better NR. When comparing Km and dHm, we observe that the use of a non-uniform lattice does not yield significant improvements in terms of PRC or NR. The size of the B_{kj} blocks can explain this behaviour. Due to the relative smaller block size comparison to the size of the spectrogram, having a non-uniform lattice does not benefit the most.

Considering w_r values, PRC and NR results are better for $w_r = 8$ than $w_r = 4$. In fact this behaviour is observed for all the studied moments. This can be related to the duration of the positive class events. Audio-cough events usually occur in bursts, the so-called cough episodes (an example is depicted in Fig. 8). A cough episode usually lasts between 500 ms and several seconds. In consequence, $w_r = 8$ better exploits the temporal dynamics of these events. As for the main drawback of the proposed method – the classification of inter-event boundary windows – it can be alleviated by using post-processing techniques to improve segmentation of the target events in such parts [51].

Both PRC and NR drop when w_c is enlarged. A coarser-grained spectral characterisation might be the underlying reason. The larger the energy patterns, the higher the number of frequency bands that must be characterised by the moments at once. On the other hand, this property allows reaching a balance between efficiency and effectiveness according to equation (5).

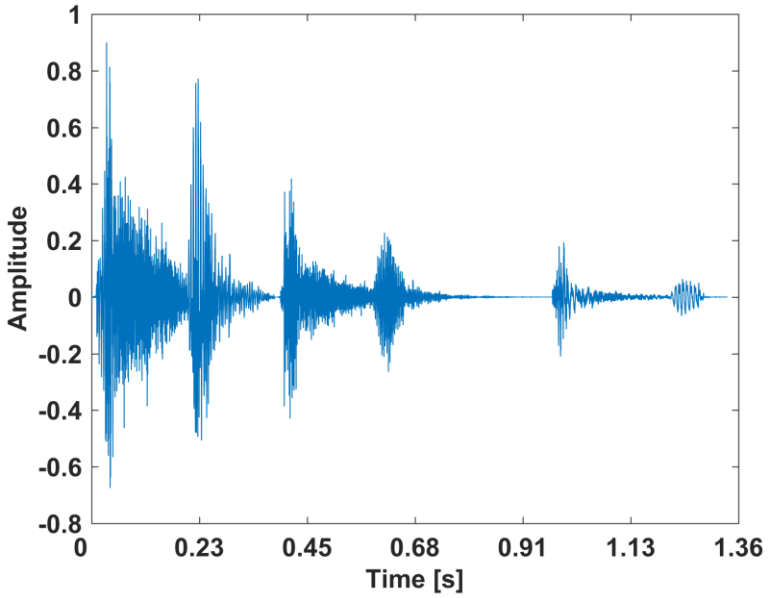


Fig. 8 Representation of a cough episode

To conclude the analysis, the feature selection step is discussed. Table 5 shows the selected features for (8 × 8) Km (similar selections were observed in other moments). The number of selected features accounts for the energy patterns for each moment order as in Table 2 – i.e $(p=0,\quad q=0)\rightarrow[1,\quad 13]$, $(1,\quad 0)\rightarrow[14,\quad 26]$, …, $(0,\quad 2)\rightarrow[66,\quad 78]$. In the light of this example, the following can be inferred: many significant features are present for every SNR value (e.g. the four most meaningful features are the same), while others are present only for two SNRs (e.g. 66 in 3 dB and 15 dB) and others are only present for one of the SNR levels (e.g. 20, 22 and 24 in -6 dB, 26 in 3 dB or 21 and 54 in 15 dB). From an overall perspective, the majority of the selected features for the three SNR values belong to combinations $(p=0,\quad q=0)$ or $(1,\quad 0)$. This supports our initial hypothesis that lower order moments are also more noise-robust in audio signal processing.

	Selected features												
SNR [dB]	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º	13º
-6	14	15	1	53	16	25	19	20	18	22	24	2	3
3	14	1	15	53	16	25	66	2	13	19	3	18	26
15	14	1	53	15	66	2	54	21	16	18	3	25	19

Table 5 Example of the selected features for each SNR; Case of Km and (8 x 8) block size

5.2 Comparison with other methods

We computed four additional feature sets to perform a comparative evaluation in Section 4.3. Results show that the proposed approach outperforms the state-of-the-art feature

sets. Additionally, differences in performance among the four compared methods can also shed some new light on the AED discipline, especially when less common audio signals – like acoustic biomedical signals – are the goal of the analysis.

The simplest method – LPCC – offers the lowest performance. On the other hand, MFCC, PNCC and SSCH exhibit different behaviours. Among these three, MFCC are the simplest and most widely applicable feature set whereas PNCC are the opposite. Besides, PNCC and SSCH were designed explicitly for robust speech detection. On this basis, it seems plausible the superiority of PNCC regarding SEN performance. PNCC incorporate a noise-suppression algorithm based on asymmetric filtering that suppresses background sounds and a module that accomplishes temporal masking [17]. They are overcome by MFCC regarding SPE though, which constitutes an interesting result. Speech signals belong to the negative class in our study while they are in the positive class according to the original design of PNCC [17]. This may be the reason why a more straightforward method such as MFCC generalises better against a diverse negative class (see Section 3.1). Other factors that may explain the PNCC behaviour are the limitation of the frequency range between 0 and 2000 Hz and the acoustic and spectral properties of audio-cough events. Speech signals cover the [0,4000] Hz range, so the PNCC noise-suppression algorithm is thought to provide its maximum capacities in such frequency range. In the same line, the acoustic and spectral properties of audio-cough events (a low-frequency signal without a clear tonal structure) are likely to make noise-suppression algorithm treat them as noise, so we do not fully benefit from this innovation.

Another remarkable result is the equivalent SEN performance between MFCC and SSCH. SSCH is a feature set based on the centroid and energy of the spectrum after the application of a filterbank. Signals like speech or music have a clear tonal structure that can be characterised through spectral centroids, spectral bandwidths or spectral crest factor measures [52]. However, other signals such as cough do not have such a clear structure. For example, the intermediate and voiced phases of an audio-cough event are very subtle, so it is difficult to consistently estimate certain properties like pitch or formants in overlapping sound conditions.

Our proposal is more complex than previous simple AED features such as MFCC or LPCC. Building the B_{kj} blocks and moment computation increase the computational load. Accuracy is not undermined by this additional complexity, although it could be argued that real-time performance is. However, taking advantage of the technological background for

image processing can compensate the extra load. Nowadays, many devices such as smartphones, laptops, or tablets have Graphical Processing Units (GPU). GPU allow parallel computation of many simple operations provided the structure of the data is suitable and that is the case for B_{kj} blocks. Hence, a GPU-implementation would increase the technology readiness level (TRL) of this proposal. Likewise, some moment algorithms have been studied in the area of applied mathematics, with currently existing methodologies for their efficient computation [45]. Finally, some moment families are actually efficient implementations of polynomial computations [41], [43], [44].

SSCH and PNCC are complex methods for speech detection that require more design choices than our proposal [17], [31]. This is a shortcoming when using them in other audio processing tasks. Conversely, our proposal is aligned with a divide-and-conquer philosophy to face AED problems. The smaller the number of parameters, the easier the method configuration. If the best-performing feature parameter configuration is not good enough, one can always resort to complementary techniques for noise suppression [53], [54], post-processing [51] or more complex classifiers [50] such as support vector machines, hidden Markov models or nearest feature line. This way, the system will offer more modularity and computing the feature set will not constitute an initial difficulty.

Finally, our performance is equivalent to previous commercial cough detectors [19], [20]. They employed other signals apart from audio-cough events, however. Alike, the method presented in [23] reported comparable SEN values whereas in [24] the recall is worse and they use several classifiers instead of only one. The SIF approach in [28] is intended for robust AED. However, it relies on a very high-dimensional feature space, which constitutes a disadvantage that limits the applicability in real-life problems.

6 Conclusions and future directions

6.1 Conclusions

This paper proposes the extension of moment theory to perform audio-cough detection. The proposal borrows the first steps from MFCC (time-frequency and application of a filter bank defined in the Mel scale) and introduces moment calculation in the latest stage to characterise energy patterns in a particular time frame and for specific frequency bands. The new feature set achieves overall good cough detection capability and noise robustness.

Our proposal is validated using a signal database with three different SNRs: -6, 3 and 15 dB. The experimental results confirm the capability of our approach to solve the AED problem at hand. Discrete orthogonal moments (Tm, Km and dHm) were the best performing. In particular, Tm offered the highest PRC and do did dHm for NR. Regarding the configuration of inner parameters, our analysis showed that they directly affect the performance of the method. Specifically, the temporal length of energy patterns (w_r) is related to the temporal dynamics of positive class events. On the other hand, the frequency length of energy patterns (w_c) partially determined the pattern recognition capabilities and noise robustness. Together with the number of filters (M), these parameters determined the dimension of the feature space. The use of lower order moments is advisable to avoid problems in the definition of the polynomials and, according to our results, they are inherently more robust against overlapping sounds.

The comparison of our approach to other methods (MFCC, LPCC, PNCC and SSCH) shows the superiority of the proposal and confirms our initial hypothesis that *ad-hoc* audio signal processing methods do not always provide the same performance when applied to other audio signals and/or in other contexts.

6.2 Future directions

The following paragraphs describe some additional future research that can be performed to seize the applicability of our proposal in AED. Firstly, in applications where the spectral content is of interest, more complex frequency decomposition methods such as parametric estimations of the PSD (e.g., Yule-Walker [32] or correntropy-based spectral characterisation [55]) could be used. On the one hand, these estimations may provide richer information at low level without affecting the main features of the method. On the other hand, their computation tends to be less efficient than in the case of the Fourier Transform via the FFT algorithm. Alternative scales can also be used for establishing the limits of the filters in the filter bank. For example, some audio processing methods use the Octave scale (e.g. the Octave spectral contrast used in music genre classification [56]) or the Bark scale as in SSCH [31]. For some other applications, changes do not need to be limited to the scale but also to the shape of the filters. Instead of triangular filters, rectangular filters or other more complex definitions such as biologically inspired gammatone filters could be used [57].

Finally, our results show the potential of the proposed methodology to become part of the *machine hearing* toolset especially for the diagnosis of diseases based on acoustic biomedical signals like cough [8], asthma wheeze [58] or lung sounds [49], which are increasingly grabbing more attention thanks to new tele-monitoring technology [59], [60], [61]. Further classification of the detected cough events can be used to differentiate between dry or productive coughs, to early detect a specific disease (e.g. lung cancer), or to assess the severity of a condition (detection of COPD –Chronic Obstructive Pulmonary Disease- exacerbations). However, not only clinical applications would benefit from the extension to moment theory here proposed. This approach could potentially be applied to any type of audio signal, since the analysis is ultimately based on the energy content in different frequency bands and does not rely on properties such as pitch or formants which can be difficult to estimate for some signals.

ACKNOWLEDGEMENTS

This work was supported by the Digital Health & Care Institute Scotland as part of the Factory Research Project SmartCough/MacMasters. Thanks are also given to Cancer Research UK for grant C59355/A22878. The authors would also like to acknowledge funding from the Centre for Excellence for Sensor and Imaging Systems –CENSIS- (project CAF-0036) and from the Royal Society of Edinburgh and National Science Foundation of China (project NNS/INT 15-16 Casaseca) which partially covered Dr. Casaseca-de-la-Higuera's, contribution. Finally, thanks are given to University of the West of Scotland for partially funding J. Monge-Álvarez's and C. Hoyos-Barceló's studentship.

REFERENCES

[1] G. A. Fontana and J. Widdicombe, "What is cough and what should be measured?," *Pulmonary Pharmacology & Therapeutics*, vol. 20, no. 4, pp. 307-312, Aug. 2007, DOI: 10.1016/j.pupt.2006.11.009

[2] A. H. Morice, *et al.*, "ERS guidelines on the assessment of cough," *European Respiratory Journal*, vol. 29, no. 6, pp. 1256-1276, Jun 2007, DOI: 10.1183/09031936.00101006

[3] M. J. Fletcher *et al.*, "COPD uncovered: an international survey on the impact of chronic obstructive pulmonary disease [COPD] on a working age population," BMC Public Health, vol. 1, no. 11, pp. 612, Aug. 2011, DOI: 10.1186/1471-2458-11-612

[4] European Respiratory Society (ERS), "The economic burden of lung disease", chapter 2 in *European Lung White Book*, pp. 16-27, 2015. Available online: http://www.erswhitebook.org/chapters/the-economic-burden-of-lung-disease/; latest visit: 24/07/2017

[5] S. A. Walke and V. R. Thool, "Differentiation nature of cough sounds in time domain analysis," 2015 International Conference on Industrial Instrumentation and Control (ICII), pp. 1022-1026, DOI: 10.1109/IIC.2015.7150896

[6] R. V. Sharan and T. J. Moir, "An overview of applications and advancements in automatic sound recognition," *Neurocomputing*, vol. 200, pp. 22-34, Aug. 2016, DOI: 10.1016/j.neucom.2016.03.020

[7] P. Dhanalakshmi, S. Palanivel and V. Ramalingam, "Pattern classification models for classifying and indexing audio signals," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 2, pp. 350-357, Mar. 2011, DOI: 10.1016/j.engappai.2010.10.011

[8] J. S. Jin, C. Xu and M. Xu, *The Era of Interactive Media*, Ed. Springer, 2013, DOI: 10.1007/978-1-4614-3501-3

[9] N. Adami, A. Cavallaro and R. Leonardi, *et al., Analysis, retrieval and delivery of multimedia content*, Ed. Springer, 2013, DOI: 10.1007/978-1-4614-3831-1

[10] T. Giannakopoulos and A. Pikrakis, "Music Information Retrieval," chapter 8 in *Introduction to audio analysis: a MATLAB approach*, Ed. Elsevier, 2014, DOI: 10.1016/B978-0-08-099388-1.00008-X

[11] A. Divakaran (Ed.), *Multimedia content analysis theory and applications*, Ed. Springer, 2009, DOI: 10.1007/978-0-387-76569-3

[12] S. Essid, G. Richard and B. David, "Musical instrument recognition by pairwise classification strategies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1401-1411, Jul. 2006, DOI: 10.1109/TSA.2005.860842

[13] J. Monge-Álvarez, C. Hoyos-Barceló, P. Lesso, *et al.*, "Effect of importance sampling on robust segmentation of audio-cough events in noisy environments," 2016 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3740-3744, DOI: 10.1109/EMBC.2016.7591541

[14] M. Rossi, S. Feese, O. Amft, *et al.*, "AmbientSense: a real-time ambient sound recognition system for smartphones," 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), pp. 230-235, DOI: 10.1109/PerComW.2013.6529487

[15] M. V. Ghiurcau, C. Rusu, R. C. Bilcu, *et al.*, "Audio based solution for detecting intruders in wild areas," *Signal Processing*, vol. 92, no. 3, pp. 829-840, Mar. 2012, DOI: 10.1016/j.sigpro.2011.10.001

[16] M. Márquez-Molina, L. P. Sánchez-Fernández, S. Suárez-Guerra, *et al.*, "Aircraft take-off noises classification based on human auditory's matched features extraction," *Applied Acoustics*, vol. 84, pp. 83-90, Oct. 2014, DOI: 10.1016/j.apacoust.2013.12.003

[17] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition," *IEEE/AMC Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1315-1329, Jul. 2016, DOI: 10.1109/TASLP.2016.2545928

[18] T. Giannakopoulos, A. Pikrakis, "Audio Datasets," Appendix C in *Introduction to audio analysis: a MATLAB approach*, Ed. Elsevier, 2014, DOI: 10.1016/B978-0-08-099388-1.00019-4

[19] M. A. Coyle *et al.*, "Evaluation of an ambulatory system for the quantification of cough frequency in patients with chronic obstructive pulmonary disease," *Cough*, vol. 4, 1:3, Aug. 2005, DOI: 10.1186/1745-9974-1-3

[20] S. J. Barry, A. D. Dane, A. H. Morice, A. D. Walmsley, "The automatic recognition and counting of cough," *Cough*, vol. 28:2:8, Sep. 2006, DOI: 10.1186/1745-9974-2-8

[21] S. Matos, S. S. Birring, I. D. Pavord, D. H. Evans, "Detection of cough signals in continuous audio recordings using hidden Markov models," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 6, Jun. 2006, DOI: 10.1109/TBME.2006.873548

[22] T. Drugman, "Using mutual information in supervised temporal event detection: application to cough detection," *Biomedical Signal Processing and Control*, vol. 10, Mar. 2014, pp. 50-57, DOI: 10.1016/j.bspc.2014.01.001

[23] M. You, *et al.*, "Novel feature extraction method for cough detection using NMF," *IET Signal Processing*, vol. 11, no. 5, pp. 515-520, Jun. 2017, DOI: 10.1049/iet-spr.2016.0341

[24] M. You, *et al.*, "Cough detection by ensembling multiple frequency subband features," *Biomedical Signal Processing and Control*, vol. 33, pp. 132-140, Mar. 2017, DOI: 10.1016/j.bspc.2016.11.005

[25] J. Amoh and K. Odame, " Deep neural networks for identifying cough sounds," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 5, pp. 1003-1011, Oct. 2016, DOI: 10.1109/TBCAS.2016.2598794

[26] P. Foggia, N. Petkov, A. Saggese, *et al.*, "Reliable detection of audio events in highly noisy environments," *Pattern Recognition Letters*, vol. 65, pp. 22-28, Nov. 2015, DOI: 10.1016/j.patrec.2015.06.026

[27] J. Dennis, H. D. Tran and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 130-133, Feb. 2011, DOI: 10.1109/LSP.2010.2100380

[28] R. V. Sharan and T. J. Moir, "Noise robust audio surveillance using reduced spectrogram image feature and one-against-all SVM," *Neurocomputing*, vol. 158, pp. 90-99, Jun. 2015, DOI: 10.1016/j.neucom.2015.02.001

[29] F. Saki and N. Kehtarnavaz, "Real-time unsupervised classification of environmental noise signals," *IEEE/ACM Transactions on Audio, Speeach, and Language Processing*, vol. 25, no. 8, pp. 1657-1667, Aug. 2017, DOI: 10.1109/TASLP.2017.2711059

[30] Y. Sun, G. Wen and J. Wang, "Weighted Spectral Features Based on Local Hu Moments for Speech Emotion Recognition," *Biomedical Signal Processing and Control*, vol. 18, pp. 80-90, Apr. 2015, DOI: 10.1016/j.bspc.2014.10.008R. J. Mammone, X. Zhang and R. P. Ramachandran, "Robust speaker recognition: a feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 58-71, Sep. 1996, DOI: 10.1109/79.536825

[31] B. Gajic and K. K. Paliwal, "Robust speech recognition in noisy environments based on subband spectral centroid histograms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 600-608, Feb. 2006, DOI: 10.1109/TSA.2005.855834

[32] S. M. Kay, *Modern Spectral Estimation: Theory and Applications*, Prentice-Hall, 1988

[33] S. X. Liao and M. Pawlak, "On image analysis by moments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 3, pp. 254-266, Mar. 1996, DOI: 10.1109/34.485554

[34] P. K. Singh, R. Sarkar, and M. Nasipuri, "A study of moment based features on handwritten digit recognition," *Applied Computational Intelligence and Soft Computing*, vol. 2016, Jan 2016, DOI: 10.1155/2016/2796863

[35] A. Broumandnia and J. Shanbehzadeh, "Fast Zernike wavelet moments for Farsi character recognition," *Image and Vision Computing*, vol. 25, no. 5, pp. 717-726, May. 2007, DOI: 10.1016/j.imavis.2006.05.014

[36] H. Zhu, H. Shu, J. Zhou, *et al.*, "Image analysis by discrete orthogonal dual Hahn moments," *Pattern Recognition Letters*, vol. 23, no. 13, pp. 1688-1704, Oct. 2007, DOI: 10.1016/j.patrec.2007.04.013

[37] H. Shu, L. Luo and J-L. Coatrieux, "Moment-based approached in imaging. Part 1, basic features," *IEEE Engineering in Medicine and Biology Magazine*, vol. 26, no. 5, pp. 70-74, Sep-Oct 2007, DOI: 10.1109/EMB.2007.906026

[38] M. E. Celebi and Y. A. Aslandogan, "A comparative study of three-moment based shape descriptors," 2005 International Conference on Information Technology: Coding and Computing (ITCC 2005), DOI: 10.1109/ITCC.2005.3

[39] M-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179-187, Feb. 1962, DOI: 10.1109/TIT.1962.1057692

[40] M. R. Teague, "Image analysis via the general theory of moment," *Journal of the Optical Society of America*, vol. 70, no. 8, 1980, DOI: 10.1364/JOSA.70.000920

[41] R. Mukundan, S. H. Ong and P. A. Lee, "Image analysis by Tchebichef moments," *IEEE Transactions on Image Processing*, vol. 10, no. 9, pp. 1357-1364, Sep. 2001, DOI: 10.1109/83.941859

[42] P-T. Yap, R. Paramesran and S-H. Ong, "Image analysis by Krawtchouk moments," *IEEE Transactions on Image Processing*, vol. 12, no. 11, pp. 1367-1377, Nov. 2003, DOI: 10.1109/TIP.2003.818019

[43] H. Zhu, H. Shu, J. Zhou, *et al.*, "Image analysis by discrete orthogonal dual Hahn moments," *Pattern Recognition Letters*, vol. 23, no. 13, pp. 1688-1704, Oct. 2007, DOI: 10.1016/j.patrec.2007.04.013

[44] K. M. Hosny, "Refined translation and scale Legendre moment invariants," *Pattern Recognition Letters*, vol. 31, no. 7, pp. 533-538, May. 2010, DOI: 10.1016/j.patrec.2009.12.008

[45] G. A. Papakostas, D. E. Koulouriotis and E. G. Karakasis, "A unified methodology for the efficient computation of discrete orthogonal image moments," *Information Sciences*, vol. 179, no. 20, pp. 3619-3633, Sep. 2009, DOI: 10.1016/j.ins.2009.06.033

[46] Universal soundbank, http://eng.universal-soundbank.com/, latest visit: 15/11/2015.

[47] J. Flusser, B. Zitova, T. Suk, *Moments and moment invariants in pattern recognition*, Ed. Wiley, Oct. 2009. ISBN: 978-0-470-69987-4

[48] I. Kononenko, E. Šimec and M. Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Applied Intelligence*, vol. 7, no. 1, pp. 39-55, Jan. 1997, DOI: 10.1023/A:1008280620621

[49] N. Sengupta, M. Sahidullah and G. Saha, "Lung sound classification using cepstral-based statistical features," *Computers in Biology and Medicine*, vol. 75, pp.118-129, Aug. 2016, DOI: 10.1016/j.compbiomed.2016.05.013

[50] T. Giannakopoulos and A. Pikrakis, "Audio Classification," chapter 5 in *Introduction to audio analysis: a MATLAB approach*, Ed. Elsevier, 2014, DOI: 10.1016/B978-0-08-099388-1.00005-4

[51] T. Giannakopoulos and A. Pikrakis, "Audio Segmentation," chapter 8 in *Introduction to audio analysis: a MATLAB approach*, Ed. Elsevier, 2014, DOI: 10.1016/B978-0-08-099388-1.00006-6

[52] A. Ramalingam and S. Krishnan, "Gaussian mixture modelling of short-time Fourier transform features for audio fingerprinting," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 4, pp. 457-463, Dec. 2006, DOI: 10.1109/TIFS.2006.885036

[53] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466-475, Sep. 2003, DOI: 10.1109/TSA.2003.811544

[54] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504-512, Jul. 2001, DOI: 10.1109/89.928915

[55] A. Garde, L. Sörnmo, R. Jané, *et al.*, "Correntropy-based spectral characterization of respiratory patterns in patients with chronic heart failure," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 8, pp. 1964-1972, Aug. 2010, DOI: 10.1109/TBME.2010.2044176

[56] C. Lee, J. Shih, K. Yu, *et al.*, "Automatic music genre classification using modulation spectral contrast feature," 2007, IEEE International Conference on Multimedia and Expo (ICME), DOI: 10.1109/ICME.2007.4284622

[57] X. Valero and F. Alias, "Gammatone cepstral coefficients: biologically inspired features for non-speech audio classification," *IEEE Transactions on Multimedia*, vol. 14, no. 6, pp. 1684-1689, Dec. 2012, DOI: 10.1109/TMM.2012.2199972

[58] M. Wiśniewski and T. P. Zieliński, "Joint Application of audio spectral envelope and tonality index in an e-asthma monitoring system," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 1009-1018, DOI: 10.1109/JBHI.2014.2352302

[59] D. Aranki, G. Kurillo, P. Yan, *et al.*, "Real-time tele-monitoring of patients with chronic heart-failure using a Smartphone: lessons learned," *IEEE Transactions on Affective Computing*, vol. 7, no. 3, pp. 206-219, Jul-Sep. 2016, DOI: 10.1109/TAFFC.2016.2554118

[60] G. Sorwar and R. Hasan, "Smart-TV based integrated e-health monitoring system with agent technology," 2012, International Conference on Advanced Information Networking and Applications Workshops (WAINA), DOI: 10.1109/WAINA.2012.155

[61] S. Mukherjee, K. Dolui and S. K. Datta, "Patient health management system using e-health monitoring architecture," 2014 IEEE International Advance Computing Conference (IACC), DOI: 10.1109/IAdCC.2014.6779357